論文審査の結果の要旨

氏名　久光　徹

本論文は6章からなり，第1章では日本語の局所構造解析，および低レベル文脈情報の定義を述べ，低レベルの文脈情報を用いた局所構造解析の提案と，論文全体の構成を述べている．「局所構造解析」とは，形態素解析，及び文節の内部構造の解析までと定義し，「低レベル」とは，「対象とする解析の出力結果に含まれる範囲の情報を越えない」ことと定義し，「文脈」とは，対象とする解析において，低レベルの情報を参照しうる一定範囲の文集合と定義している．執筆者が提案している方法は，「低レベルの文脈情報」を必要に応じて解析対象の文の周囲の「文脈」から獲得し，これを従来型の手法と組み合わせて解析精度を向上させようとする方法である．従来の枠組みでは，大量の正解データや高度な意味的情報等を用いない限り，局所構造解析の一層の精度向上は困難とされており，頑健性や，正解データ獲得の点で困難に直面していた．

第2章では，局所構造解析の基礎となる形態素解析について，解析アルゴリズム，コスト関数の設定方法，動詞活用処理の最適化について述べている．解析アルゴリズムは，日本語形態素解析の諸手法を一般的に取り扱うために論文執筆者が提唱した，「接続コスト最小法」と呼ぶ枠組に沿って，*N-best* 解を導出するためのアルゴリズムが述べられており，これを利用して解析の曖昧性を縮約・保持し，文脈情報の情報を利用することを可能としている．また，接続コスト最小法の枠組みを用いたヒューリスティク・コスト関数の体系的な設定方法について論じており，「一般化文節数最小法」の精度を向上させる手法について具体的に述べている．同章ではさらに，動詞の活用を処理するための辞書見出しの最適化についても論じており，音韻論的分析の合理性を保ちつつ，日本語の漢字仮名混じり表記に適合した処理方式を提案し，形態素解析の効率と頑健性を向上できることを示している．

3章では，2章で述べた形態素解析に基づき，文脈情報の古典的な利用形態の一つである，パターンマッチによる書き換え規則を用いた形態素解析の後処理方法について述べ，「窓」と呼ぶ文集合内で低レベル文脈情報を参照することにより，後処理の精度を一層向上させる方法について述べている．書き換えルールは，複雑な誤りパターンをもつ日本語形態素解析に適用するために新たに考案した，誤り駆動型の自動学習手法を用いて獲得し，これに加えて，「窓」と呼ぶ複数の文集合（例として1記事を用いる場合が示されている）に対する解析結果をプールし，その中での書き換え結果を参照することにより，書き換え規則だけを用いた場合に比べ，曖昧性の解消や未登録語の同定の能力がより一層向上することを示している．これらの後処理により，未知語を含む条件下で，recall，precision ともに 2〜3%程度の向上が可能であることが示

されている．

第４章では，文書走査法と呼ぶ手法による，漢字で書かれた名詞連鎖による複合名詞の，名詞間の係り受け構造の解析方法について述べている．複合名詞の構造を規定するルールとヒューリスティクスに加え，複合名詞を構成する単語の共起情報を文脈中から獲得することにより，高精度かつ未登録語に対して頑健な構造解析が可能であることを示している．名詞間の共起は，定められたテンプレートの二つの変数部分にマッチすることとして定義し，従来の共起の定義に比べ，獲得される共起情報の質と量が向上することが示されている．また，形態素解析において解析に失敗した未知語を，共起情報の獲得中にテンプレートにより同定し，共起情報抽出の対象に繰り入れることにより，未知語に対しても頑健な複合名詞解析が実現できるとしている．解析の精度は，3 単語相当で 88%であり，未知語の存在を考慮すれば，従来手法を 20%以上上回ることがしめされており、 画期的な結果となっている．

第５章では，ルールと統計情報を併用した未知語獲得について述べている．３，４章で示された方法では解析時に同定が困難な，人名，社名等の固有名詞や，略称等を対象とし，単語同定のための少数のルールと，文書全体から得られる簡単な統計量を組み合わせて，新聞紙上に現れる未知語の 40%程度を， 95%前後の精度で収集できることを示しているおり、 今後の言語処理手法の基盤を与えている．

なお，本論文第２章，第４章は新田義彦氏との共同研究，第３章，第５章は，丹羽芳樹氏との共同研究に基づいているが，論文提出者が主体となって分析，及び検証を行ったもので，論文提出者の寄与が充分であると判断する．したがって，博士（理学）を授与できると認める．